Genetic Variation and Relationships among Cultivated, Wild, and Semiwild Soybean

Yiwu Chen and Randall L. Nelson*

ABSTRACT

Some annual *Glycine* accessions are intermediate between the standard phenotypes of *Glycine max* (L.) Merr. and *Glycine soja* Sieb. & Zucc. and have been labeled semiwild. Few studies have examined both the genetic and phenotypic relationships among *G. soja*, *G. max*, and semiwild-types by combining morphological traits and DNA markers. The objectives of this research were to quantify genetic variation within *G. soja*, *G. max*, and semiwild accessions; to investigate the relationships among the *G. soja*, *G. max*, and semiwild accessions; and to examine the relationships among phenotypes on the basis of morphological traits and genotypes on the basis of DNA markers. Ninety-two semiwild, *G. soja*, and *G. max* accessions from the USDA Soybean Germplasm Collection were evaluated for 20 phenotypic traits and with 137 RAPD markers. Mahalanobis distances and a Jaccard genetic similarity matrix were calculated for phenotypic traits and DNA data, respectively. Nonhierarchical and hierarchical clustering as well as multidimensional scaling (MDS) were used to evaluate relationships among semiwild, *G. soja*, and *G. max* accessions. Principal component analysis was applied to identify the morphological traits that were most significant in separating the three groups. For the accessions examined, unique RAPD markers were found for each taxonomic type. Three clusters defined by either phenotypic or DNA data are highly consistent and strongly corresponded to *G. soja*, *G. max*, and semiwild classifications. On the basis of the analysis of RAPD data, *G. soja* accessions have the greatest genetic diversity and semiwild accessions the least. *Glycine max* and semiwild accessions are more closely related to each other than to *G. soja* accessions. These data will be useful in helping to define a core collection of annual *Glycine.*

There are two species usually recognized within the genus *Glycine* subgenus *Soja, Glycine max* and *Glycine soja*. On the basis of data from morphology (Palmer et al., 1987), cytogenetics (Hymowitz and Singh, 1987), phytoalexins (Keen et al., 1986), restriction endonuclease fragment analysis of mitochondrial DNA (Doyle, 1988), ribosomal RNA (Doyle and Beachy, 1985), chloroplast DNA (Shoemaker et al., 1986), and sequences from the ITS region of nuclear ribosomal DNA (Kollipara et al., 1997), *G. soja* is considered the ancestor of *G. max*. Besides *G. max* and *G. soja*, an intermediate form sometimes known as *G. gracilis* Skvortz. has been described. This form has numerous characteristics intermediate between *G. max* and *G. soja* and was first proposed as a new species by Skvortzow (1927). Fukuda (1933) proposed that *G. gracilis* is an intermediate evolutionary type between *G. soja* to *G. max*, but Hymowitz (1970) suggested that *G. gracilis* is a hybridization product of *G. max* and *G. soja*. The latter hypothesis was supported by Broich and Palmer (1981) on the basis of the results from their study of the frequency and distribution of 10 alleles among *G. max*, *G. soja*, and *G. gracilis* accessions. On the basis of numerical taxonomic analysis, Broich and Palmer (1980, 1981) recommended that the designations *G. max* and *G. gracilis* both be utilized. They reasoned that *G. gracilis* phenotypes can be distinguished from *G. max* and they represent an intermediate form of domesticated soybean. In addition, *G. max* and its semiwild relative should be regarded as taxonomically distinct from *G. soja* since both are domesticated.

Hermann (1962) removed *G. gracilis* from the species rank and incorporated it into *G. max* on the basis of classical taxonomy. Several studies support the elimination of *G. gracilis* as a separate species. Singh and Hymowitz (1989) demonstrated that *G. max*, *G. soja*, and *G. gracilis* all hybridized readily, and F_1 seeds produced viable, vigorous, and fertile plants with normal meiotic pairing. Wang (1976) suggested all three classifications in subgenus *Soja* could be a single species since they were not reproductively isolated, but on the basis of cultivated status, he recommended that *G. soja* be kept as a species and *G. gracilis* reclassified as *G. max*.

Dae et al. (1995) applied isozyme and RAPD techniques to evaluate genetic variation within the subgenus *Soja* and concluded on the basis of morphological appearances that the intermediate forms of *G. max* were also intermediate between *G. max* and *G. soja* on the basis of genotypic measurements. Fei and Chen (1996) analyzed genetic diversity of the *Glycine* genus with RAPD markers using 21 accessions from 10 species of the *Glycine* subgenus and the three species of the *Soja* subgenus (*G. max*, *G. gracilis*, and *G. soja*) with eight primers. In this analysis, they found that the three species within the S*oja* subgenus were clustered as one group with *G. gracilis* classified a subgroup within *G. max.* This research supports the idea that there should only be one species, and earlier Smartt (1984) had proposed that *G. max*, *G. soja*, and *G. gracilis* should all be classified as subspecies. Although there are many arguments about the designations of species in subgenus *Soja,* it is well accepted that *G. soja* is the ancestor of cultivated soybean and most taxonomists have kept *G. max* and *G. soja* as separate species.

Several traits have distinct differences between *G. max* and *G. soja* accessions. In general, *G. soja* has much smaller seeds (<3.0 g 100 seeds^{-1}) than *G. max*

Y. Chen, Dep. of Crop Sciences, 1101 W. Peabody Dr., University of Illinois, Urbana, IL 61801; R.L. Nelson, USDA-ARS, Soybean/Maize Germplasm, Pathology, and Genetics Research Unit, Dep. of Crop Sciences, 1101 W. Peabody Dr., University of Illinois, Urbana, IL 61801. Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the USDA or the University of Illinois and does not imply its approval to the exclusion of other products or vendors that may also be suitable. Received 16 Dec. 2002. *Corresponding author (rlnelson@uiuc.edu).

Published in Crop Sci. 44:316–325 (2004).
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: MDS, multidimensional scaling; MG, maturity group; PCA, principal component analysis; PCR, polymerase chain reaction; PI, Plant Introduction; RAPD, random amplified polymorphic DNA.

(generally >9.0 g 100 seeds^{-1}). *Glycine soja* also has viney and twining stems, severe shattering before plant maturity, and impermeable seed coats, which are all rare in *G. max*. *Glycine soja* also has much lower oil and oleic acid concentration, and higher linolenic acid concentration. There are many accessions in annual *Glycine* collections that are intermediate between the typical *G. soja* and *G. max* types. Chang et al. (1999) reported that among the 17 613 accessions of the Chinese *G. max* collection 1.5% of the accessions have 100-seed weights of less than 6.0 g, and 33% of the accessions were between 6.1 and 12.0 g 100 seeds^{-1}. Dong et al. (1999) examined 6172 *G. soja* accessions in the Chinese wild soybean collection and found that 8.5% of the accessions have 100-seed weights of more than 5.0 g. The appropriate classification of these intermediate types is not well defined. The objectives of this research were to quantify genetic variation within the *G. soja, G. max*, and semiwild *Glycine* accessions; to investigate the relationships among the *G. soja*, *G. max*, and semiwild accessions; and to examine the relationships among phenotypes on the basis of morphological traits and genotypes on the basis of DNA markers.

MATERIALS AND METHODS

Thirty semiwild, 31 *G. max,* and 31 *G. soja* accessions, previously classified on the basis of morphological traits when the accessions were initially evaluated, were selected from the USDA Soybean Germplasm Collection for this study. Accessions within each group were selected to have similar origins and maturity dates (Table 1). All *G. max* accessions are primitive types that predate scientific plant breeding. The lines were evaluated at Urbana, IL, in 1999 and 2000. The *G. max* and semiwild accessions were grown in three replications in one-row plots 2.5 m long and 0.75 m apart, and the *G. soja* lines were grown in hill plots 0.75 m apart in an aphid-proof cage with one replication in 1999. In 2000, the experiment was repeated with three replications for all entries.

Twenty phenotypic characters were selected to evaluate the differences among the groups. Eight descriptive traits included flower, pubescence, pod, seed coat, and hilum color; pubescence form; pubescence density; and seed coat luster. Agronomic data consisted of a lodging score (scored 1 = erect to 5 = prostrate), a shattering score (scored at harvest and 2 wk after maturity with the following scale: 1 = no shattering, 2 = 1 to 10% 3 = 11 to 25% 4 = 26 to 50% 5 = over 50%), weight 100 seeds^{-1}, the ratio of stem diameter at the first internode and the last internode measured on three plants per plot, and terminal leaflet shape. Terminal leaflet shape was based on the ratio of the maximum length of the leaflet by the maximum width of the leaflet on three plants in each plot. The sample leaflets were taken at approximately two-thirds of the distance from ground to the top of the final plant height. Stem diameter and leaflet measurements were made late in the R6 growth stage.

Seed composition measurements included protein and oil concentration, and concentration of the following fatty acids: palmitic, stearic, oleic, linoleic, and linolenic. Nitrogen content of whole seed was determined with a LECO FP-428 Nitrogen Determinator (LECO Corp., St. Joseph, MI). The 6.25 conversion factor was used to calculate protein concentration on a dry weight basis. Oil concentration (dry weight basis) of whole seed was determined with a 5 MHz nuclear magnetic resonance spectrometer (Newport Oxford Instruments, Newport Pagnell, England). Fatty acid methyl esters were prepared from chloroform/hexane/methanol (8:5:2, v/v/v) extracts of crushed seed by transmethylation with sodium methoxide. Fatty acid composition was determined with a Hewlett-Packard 5890-II (Palo Alto, CA) gas chromatograph equipped with dual flame ionization detectors, and a 0.53-mm by 30-m AT-Silar capillary column (Alltech Associates, Deerfield, IL). Authentic fatty acids were used for calibration.

Genomic DNA was isolated from the first trifoliate leaves of five greenhouse grown seedlings for each accession. Harvested leaves were placed in 15-mL screw-cap tubes and frozen at –80°C before lyophilizing the tissue. Four glass beads were added to each tube and shaken on a shaker for 3 min. DNA was extracted by the CTAB (hexadecyltrimethyl ammonium bromide) method of Kisha et al. (1997). The DNA concentration of all extracted samples was calculated from spectrophotometer readings at wavelengths of 260/280 and adjusted to a concentration of 10 ng μL^{-1}. Forty-four decanucleotide primers from Operon Technologies Inc. (Alameda, CA) were chosen for this study (Table 2). These included 35 primers of a core set identified by Thompson and Nelson (1998) and nine randomly selected primers. The amplification protocol of Kresovich et al. (1994) was used with minor modifications. Amplified products were separated by 1% (w/v) agarose gels in 1× Tris-acetate buffer for 2.5 h at 125 V with constant power, stained with ethidium bromide, and visualized under UV light.

A Mahalanobis distance matrix was calculated for 12 quantitative traits collected in 1999 and 2000 by the formula: $D^2\ (i/j) = (\overline{X}_i - \overline{X}_j)'COV^{-1}\ (\overline{X}_i - \overline{X}_j)$ and PROC DISCRIM Mahalanobis in PC SAS (SAS Institute, 1999). In this formula COV^{-1} is the inverse of the pooled sample variance-covariance matrix, and $\overline{X}_i$ and $\overline{X}_j$ are the respective vectors of measurements on groups i and j. Principal component analysis was employed to identify the main factors among the 12 measured characters. Variables in this study were not measured in the same units, so the data were standardized with a square root transformation. The standardized data were subjected to principal component analysis by PROC PRINCOMP and VARCLUS option of PROC CLUSTER in PC SAS (SAS Institute, 1999).

RAPD fragments were scored as either present (1) or absent (0). Jaccard's coefficient was used to measure the distance between each pair of genotypes with the following formula: $S_{ij} = a/(a + b + c)$, where a is the number of common bands; b is the number of bands present in first accession and absent in the second; and c is the number of bands absent in first accession and present in the second. $D_{ij} = 1 - S_{ij}$ was calculated as a measure of dissimilarity.

A hierarchical cluster analysis was performed on the 92 by 92 genetic dissimilarity matrix using the WARD option of PROC CLUSTER of PC SAS (SAS Institute, 1999). Mean distances within and between clusters were calculated using a SAS Interactive Matrix Language (SAS/IML, SAS Institute, 1999) program provided by D.Z. Skinner (personal communication, 2001). Values of the cubic clustering criterion (CCC), pseudo F statistic (PSF), and Hotelling's pseudo T^2 statistic were also considered for defining optimum cluster numbers (SAS Institute, 1999). A nonhierarchical cluster analysis procedure, VARCLUS option of PROC CLUSTER in PC SAS (SAS Institute, 1999), was also applied to the original fragment data to divide the accessions into nonoverlapping clusters. The data were also subjected to principal component analysis.

The matrix of genetic distances generated from Jaccard's genetic dissimilarity coefficient was subjected to multidimensional scaling (MDS) (Shepard, 1974) by the MDS procedure in PC SAS (SAS Institute, 1999). The ABSOLUTE option

Table 1. *Glycine max*, *G. soja*, and semiwild accessions used for phenotypic evaluation and RAPD analysis.

Code	Class	PI number	Province or area	Country	MG
G01	Semiwild	PI 417139	Tohoku	Japan	I
G06	Semiwild	PI 416762	Tohoku	Japan	II
G08	Semiwild	PI 65388	Heilongjiang	China	II
G10	Semiwild	PI 232992	Fukui	Japan	III
G11	Semiwild	PI 232987	Northeast	China	II
G12	Semiwild	PI 468919	Liaoning	China	III
G13	Semiwild	PI 437662	Jilin	China	II
G16	Semiwild	PI 476938	Northern	Vietnam	III
G23	Semiwild	PI 232989	Northeast	China	II
G25	Semiwild	PI 417138	Tohoku	Japan	II
G28	Semiwild	PI 437918	Unknown	China	I
G31	Semiwild	PI 81771	Northeast	China	II
G34	Semiwild	PI 86046	Hokkaido	Japan	II
G37	Semiwild	PI 253651C	Unknown	China	III
G39	Semiwild	PI 291309C	Heilongjiang	China	I
G45	Semiwild	PI 81763	Northeast	China	II
G46	Semiwild	PI 291275	Heilongjiang	China	I
G49	Semiwild	PI 291277	Heilongjiang	China	I
G59	Semiwild	PI 438152	Primorye	Russia	II
G61	Semiwild	PI 79593	Heilongjiang	China	II
G63	Semiwild	PI 468907	Jilin	China	I
G66	Semiwild	PI 81772	Northeast	China	I
G67	Semiwild	PI 483459	Jilin	China	I
G69	Semiwild	PI 135590	Heilongjiang	China	II
G70	Semiwild	PI 437944	Unknown	Russia	II
G75	Semiwild	PI 461509	Jilin	China	I
G76	Semiwild	PI 79648	Liaoning	China	I
G84	Semiwild	PI 437116	Far East	Russia	I
G88	Semiwild	PI 79727	Heilongjiang	China	I
G90	Semiwild	PI 326580	Unknown	Germany	I
M02	*G. max*	PI 68765	Northeast	China	II
M03	*G. max*	PI 88810	Pyongan Puk	Korea, North	II
M04	*G. max*	PI 86741	Northeast	China	II
M07	*G. max*	PI 79756	Heilongjiang	China	II
M14	*G. max*	PI 54854	Northeast	China	I
M17	*G. max*	PI 79692	Heilongjiang	China	III
M19	*G. max*	PI 88282	Jilin	China	III
M24	*G. max*	PI 88797	Northeast	China	I
M26	*G. max*	PI 79699	Heilongjiang	China	I
M27	*G. max*	PI 437493	Primoreye	Russia	II
M30	*G. max*	PI 88997	Northeast	China	II
M32	*G. max*	PI 417076	Tohoku	Japan	I
M33	*G. max*	PI 89003-1	Northeast	China	II
M35	*G. max*	PI 92569	Jilin	China	II
M36	*G. max*	PI 91110-1	Heilongjiang	China	I
M38	*G. max*	PI 91119	Heilongjiang	China	II
M40	*G. max*	PI 30594	Heilongjiang	China	II
M42	*G. max*	PI 70027	Heilongjiang	China	I
M43	*G. max*	PI 232993	Fukui	Japan	II
M44	*G. max*	PI 96195	Liaoning	China	II
M53	*G. max*	PI 89138	Hamgyong Puk	Korea, North	II
M55	*G. max*	PI 68474-2	Northeast	China	I
M74	*G. max*	PI 437119	Primorye	Russia	I
M77	*G. max*	PI 79609	Heilongjiang	China	II
M79	*G. max*	PI 68572	Heilongjiang	China	I
M80	*G. max*	PI 68475-1	Northeast	China	II
M82	*G. max*	PI 92698	Jilin	China	II
M85	*G. max*	PI 476911	Northern	Vietnam	II
M86	*G. max*	PI 68728	Northeast	China	II
M89	*G. max*	PI 437101	Far East	Russia	I
M92	*G. max*	PI 437476	Primorye	Russia	III
S05	*G. soja*	PI 483460B	Liaoning	China	III
S09	*G. soja*	PI 464891B	Jilin	China	II
S15	*G. soja*	PI 464890A	Jilin	China	II
S18	*G. soja*	PI 479753B	Jilin	China	II
S20	*G. soja*	PI 101404B	Heilongjiang	China	II
S21	*G. soja*	PI 424004B	Kyonggi	Korea, South	II
S22	*G. soja*	PI 407288	Jilin	China	II
S29	*G. soja*	PI 424004A	Kyonggi	Korea, South	II
S41	*G. soja*	PI 342618B	Primorye	Russia	I
S47	*G. soja*	PI 79752	Jilin	China	I
S48	*G. soja*	PI 479749	Jilin	China	III
S50	*G. soja*	PI 407297	Liaoning	China	II
S51	*G. soja*	PI 479748	Jilin	China	II
S52	*G. soja*	PI 342620A	Primorye	Russia	I
S54	*G. soja*	PI 406684	Hokkaido	Japan	III
S56	*G. soja*	PI 81762	Amur	Russia	II
S57	*G. soja*	PI 514674	Hokkaido	Japan	III

Continued on next page.

Table 1. Continued.

Code	Class	PI number	Province or area	Country	MG
S58	*G. soja*	PI 479750	Jilin	China	I
S60	*G. soja*	PI 407296	Liaoning	China	II
S62	*G. soja*	PI 342622A	Primorye	Russia	I
S64	*G. soja*	PI 522182B	Heilongjiang	China	I
S65	*G. soja*	PI 468916	Liaoning	China	III
S68	*G. soja*	PI 464891C	Jilin	China	II
S71	*G. soja*	PI 507581	Aomori	Japan	III
S72	*G. soja*	PI 407298	Liaoning	China	II
S73	*G. soja*	PI 407299	Liaoning	China	II
S78	*G. soja*	PI 440913A	Jilin	China	II
S81	*G. soja*	PI 479747	Jilin	China	III
S83	*G. soja*	PI 479746B	Jilin	China	II
S87	*G. soja*	PI 407289	Jilin	China	II
S91	*G. soja*	PI 479744	Jilin	China	I

was used to maintain the scale of 0 and 1 for making interpretation and graphing easier. The criteria are similar to that described by Thompson et al. (1998) and Gizlice et al. (1996). To evaluate the effectiveness of 2 to 22 dimensions, the goodness of fit criterion (R^2) between the original data and the predicted values that were derived from the MDS coordinates was used. The best MDS analysis was considered to be the fewest dimensions that resulted an $R^2 > 0.95$ with the original genetic distance matrix. The matrix of the Mahalanobis distances from twelve phenotypic traits was also subjected to multidimensional scaling.

RESULTS AND DISCUSSION

Variation Based on Phenotypic Data

On the basis of phenotypic data, *G. max* has the greatest diversity and *G. soja* has the least. All of the evaluated qualitative traits are uniform for *G. soja* except for pubescence form. Purple flowers, tawny pubescence color, normal pubescence density, seed coat bloom, and black pod, seed coat, and hilum color are common for all *G. soja* entries. Twelve quantitative traits (lodging, shattering, leaflet shape, stem ratio, seed weight, and seed concentrations of protein, oil, and five fatty acids) were subjected to analysis of variance (Table 3). Statistically significant differences were found between years for all traits except stem ratio, protein, and oleic acid concentration; however, the differences between the years were quite small for most traits. The means of the three taxonomic classes for all 12 traits were nearly all significantly different, but most of the traits have overlapping ranges across the three classes (Table 3). Seed weight, oil concentration, oleic, and linolenic acid concentration have highly significant differences among the three classes and little or no overlap in ranges of values (Table 3). *Glycine soja* has a viney stem that is never erect, severe shattering, a small stem ratio (<4.5), low seed weight (<2.5 g 100 seeds^{-1}), low oil concentration (<130 mg g^{-1}), low oleic acid concentration (<140 mg g^{-1}), and high linolenic acid concentration (>140 mg g^{-1}) (Table 3). G*lycine max* is variable for lodging and shattering, has a high stem ratio (>4.5), larger seed weight (>9.0 g 100 seeds^{-1}), high oil concentration (>185 mg g^{-1}), high oleic acid concentration (>190 mg g^{-1}), and low linolenic concentration (<95 mg g^{-1}). The semiwild accessions are intermediate between *G. soja* and *G. max* for most traits (Table 3).

On the basis of the Mahalanobis distance matrix calculated from the data collected in 2000, the Ward's method assigned all accessions into three clusters, which corresponded closely to the original accession classifications. Cluster 1 is composed of 31 *G. max* accessions and four semiwild lines (G75, G16, G70, and G06); cluster 2 has all 31 *G. soja* entries; and cluster 3 contains 26 semiwild accessions. The four semiwild exceptions

Table 2. The sequences of 44 primers used to characterize the genetic diversity of 92 *G. max*, *G. soja*, and semiwild accessions and the number of fragments produced.

Primers	Sequence 5→3′	Total number of fragments	Number of polymorphic fragments
OPA-20	AATCGGGCTG	5	4
OPE-01	CCCAAGGTCC	7	7
OPF-04	GGTGATCAGG	9	3
OPG-04	AGCGTGTCTG	13	10
OPG-06	GTGCCTAACC	5	4
OPG-11	TGCCCGTCGT	7	5
OPH-02	TCGGACGTGA	8	7
OPH-12	ACGCGCATGT	2	1
OPH-13	GACGCCACAC	2	2
OPH-15	AATGGCGCAG	8	0
OPK-01	CATTCGAGCC	7	6
OPK-03	CCAGCTTAGG	5	4
OPK-10	GTGCAACGTG	7	2
OPK-16	GAGCGTCGAA	2	2
OPL-04	GACTGCACAC	2	1
OPL-09	TGCGAGAGTC	7	5
OPL-18	ACCACCCACC	9	6
OPM-18	CACCATCCGT	3	3
OPN-03	GGTACTCCCC	5	5
OPN-08	ACCTCAGCTC	5	0
OPN-09	TGCCGGCTTG	4	0
OPN-18	GGTGAGGTCA	5	4
OPO-01	GGCACGTAAG	13	11
OPO-04	AAGTCCGCTC	5	4
OPO-05	CCCAGTCACT	15	7
OPO-08	CCTCCAGTGT	4	2
OPO-14	AGCATGGCTC	7	2
OPO-19	GGTGCACGTT	8	3
OPP-07	GTCCATGCCA	8	2
OPP-09	GTGGTCCGCA	4	0
OPP-10	TCCCGCCTAC	9	8
OPP-11	AACGCGTCGG	4	0
OPQ-08	CTCCAGCGGA	2	0
OPR-07	ACTGGCTTGA	10	0
OPR-10	CCATTCCCCA	8	5
OPR-12	ACAGGTGCGT	11	0
OPR-13	GGACGACAAG	6	1
OPS-01	CTACTGCGCT	10	0
OPS-03	CAGAGGTCCC	9	2
OPS-05	TTTGGGGCCT	7	1
OPS-11	AGTCGGGTGG	8	0
OPS-14	AAAGGGGTCC	5	4
OPV-08	GGACGGCGTT	5	0
OPX-05	CCGCTACCGA	4	3
Total		231	137

Table 3. Accession ranges and class means for phenotypic data collected in 1999 and 2000 for three taxonomic classes of annual *Glycine*.

Trait	Class	Range of accession means	Class mean
Lodging (score of 1 to 5)	*G. max*	1 to 4	2.4 a†
	Semiwild	2 to 4	3.6 b
	G. soja	5	5.0 c
Shattering (score of 1 to 5)	*G. max*	1 to 5	2.5 a
	Semiwild	3 to 5	4.1 b
	G. soja	5	5.0 c
Leaflet shape (length/width)	*G. max*	1.9 to 2.6	2.2 a
	Semiwild	1.9 to 2.6	2.1 a
	G. soja	2.1 to 4.5	2.8 b
Stem ratio‡	*G. max*	4.6 to 8.9	6.5 a
	Semiwild	2.4 to 9.7	6.0 b
	G. soja	2.5 to 4.3	3.4 c
Seed weight (g 100 seeds^{-1})	*G. max*	8.7 to 16.8	13.1 a
	Semiwild	2.9 to 8.3	5.5 b
	G. soja	1.0 to 2.3	1.4 c
Protein (mg g^{-1})	*G. max*	366 to 429	401 a
	Semiwild	386 to 457	418 b
	G. soja	418 to 506	465 c
Oil (mg g^{-1})	*G. max*	185 to 216	200 a
	Semiwild	136 to 195	154 b
	G. soja	96 to 123	107 c
Palmitic acid (mg g^{-1})	*G. max*	96 to 127	116 a
	Semiwild	112 to 134	125 b
	G. soja	106 to 126	114 c
Stearic acid (mg g^{-1})	*G. max*	37 to 55	41 a
	Semiwild	34 to 46	41 a
	G. soja	32 to 39	34 b
Oleic acid (mg g^{-1})	*G. max*	190 to 293	234 a
	Semiwild	162 to 231	184 b
	G. soja	97 to 142	116 c
Linoleic acid (mg g^{-1})	*G. max*	470 to 561	528 a
	Semiwild	523 to 574	543 b
	G. soja	537 to 591	559 c
Linolenic acid (mg g^{-1})	*G. max*	59 to 95	81 a
	Semiwild	82 to 122	107 b
	G. soja	145 to 207	177 b

† Means with the same letter are not significantly different (p = 0.01) based on T test.

‡ Ratio of stem diameter for the first internode and the last internode on the main stem.

within the predominant *G. max* cluster 1 are all in same subcluster. All four accessions have seed weights greater than 8 g 100 seeds^{-1}, oil concentrations greater than 170 mg g^{-1}, and linolenic acid concentrations of 100 mg g^{-1} or less. G75, G70, and G06 also have high oleic acid concentrations (190–220 mg g^{-1}), and large stem ratios (7.4–9.7). These values are more typical for *G. max.* G16 has severe lodging and shattering, and a smaller than average stem ratio that is more typical of wild and semiwild accessions. The analysis of the data collected in both 1999 and 2000 resulted in the same major clusters.

Five dimensions in multidimensional scaling adequately captured the information in the original Mahalanobis distance matrix (R^2 = 0.97). Data from both years resulted in nearly identical MDS plots and results were consistent with the Ward's clustering method. The first dimensions accounted for 59% of the total variation and the two-dimensional plot showed that *G. max* and *G. soja* were in two distinct groups with the semiwild accessions generally distributed between these two groups (Fig. 1). The majority of semiwild accessions were clearly separated from the two species, but G75, G06, G70, and G16, the semiwild accessions in the predominantly *G. max* cluster by Ward's method, are positioned in the *G. max* group (Fig. 1). In the fifth dimension of the MDS, it was possible to separate G06 from the *G. max* group. Two other lines (G10 and G45) were on the boundary between the semiwild and *G. max* groups, and it is difficult to define them as either *G. max* or semiwild on the basis of this analysis, but in the fourth dimension G10 could be distinguished from the G. max accessions. The *G. soja* accessions were more tightly clustered than the other groups indicating less variability for these phenotypic traits (Fig. 1).

The results from the principal component analysis were similar for both years with the first principal component accounting for 85% of the variation in 2000. The principal component plot is similar to the MDS plot in terms of the distribution of each of the three types. G70, G06, and G16 are all near the *G. max* group and G10, G28, G88, and G66 were located distant from the other semiwild accessions in both years. Except for G10, they have similar origin and maturity. G75, which was associated with the *G. max* accessions in the previous two analyses, was removed from the *G. max* accessions in this plot but was also separated for the other semiwild accessions. S21 and S29 were the only two lines separated from the tight cluster of *G. soja* accessions and both originated from Kyonggi, South Korea, and are in maturity group II. Five of the 12 measured traits were defined by the first principal component score as significant factors. Oil concentration and seed weight are highly correlated (R value $>$ 0.9), so we included only seed weight along with stem ratio, oleic, and linolenic acid concentration as the four traits that make the most significant contributions to the total variance. The ratio of stem diameter at the top and bottom of the plant was measured for the first time in this study. This ratio, like all of the other three traits, clearly separated *G. max* and *G. soja*, but not semiwild soybean from the two species (Table 3).

The VARCLUS analysis resulted in four clusters with both 1999 and 2000 data. Cluster 1 and cluster 2 included 29 semiwild accessions. Although half of the accessions in cluster 1 weigh less than 4.0 g 100 seeds^{-1}, the mean 100-seed weight of cluster 1 (6.3 g) is nearly the same as cluster 2 (5.8 g). The range of seed weights in cluster 2 is from 4.4 to 8.9 g 100 seeds^{-1}. Cluster 3 contained all of the 31 *G. max* lines and one semiwild accession (G06) with a 12.9 g 100-seed weight mean. Cluster 4 was composed of all 31 *G. soja* lines and had the smallest 100-seed weight (mean = 1.4 g) compared with other three clusters. All four analytical procedures (Ward's, MDS, PCA, and VARCLUS) identified G06 as not being part of the semiwild group. G75, G70, and G16 were identified as such by all but the VARCLUS procedure. From the phenotypic data, we can conclude that the *G. max* and *G. soja* groups are clearly distinct from each other. Those classified as semiwild form an intermediate but not always unambiguous grouping.

Genetic Relationships Based RAPD Profiles

Forty-four primers generated 137 polymorphic fragments out of a total of 231 fragments (Table 2). The percentage of polymorphism (59%) is higher than reported

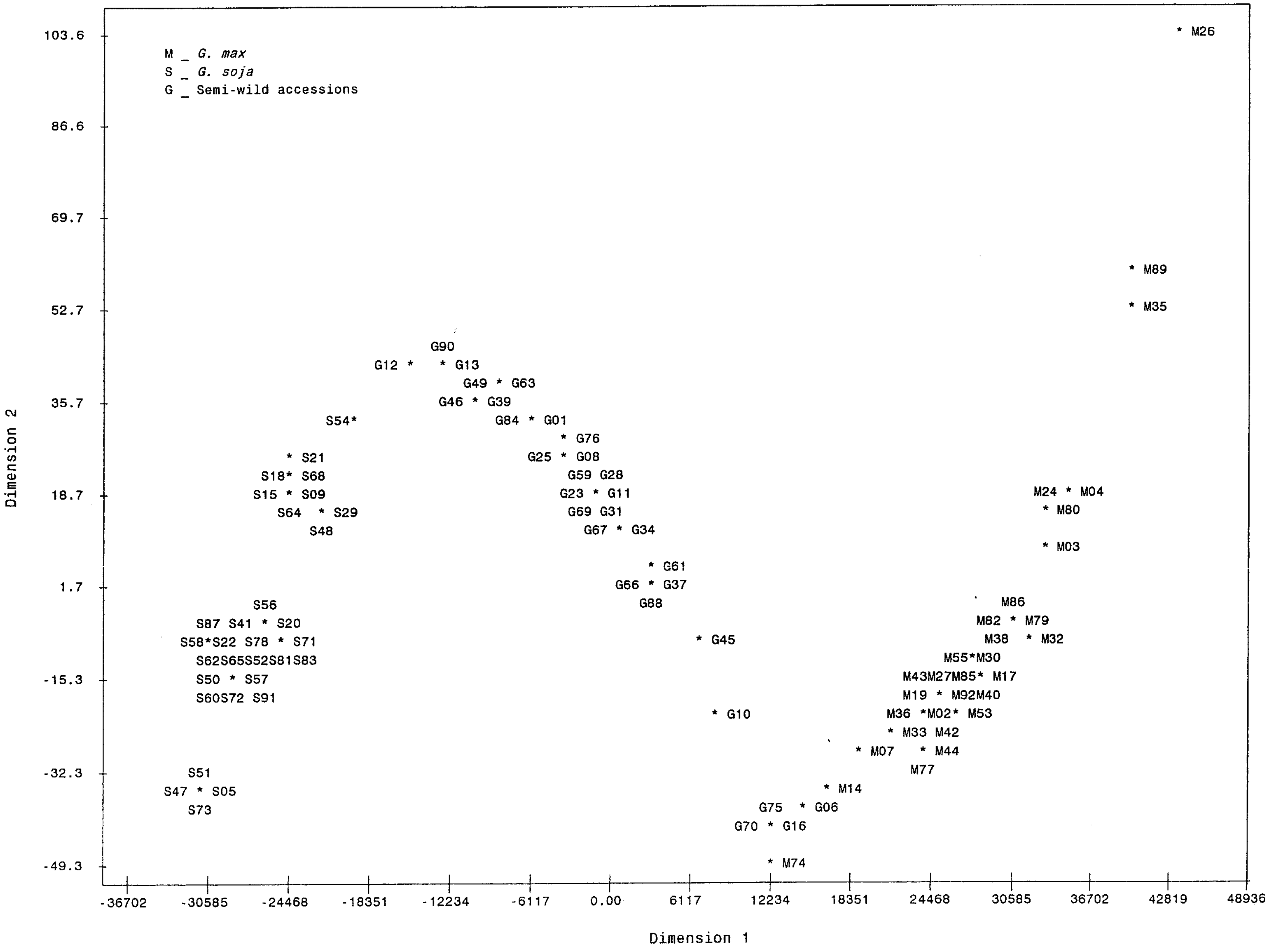

Fig. 1. Two dimensional scatter plot of 92 accessions of *G. max*, *G. soja*, and semiwild accessions obtained from multidimensional scaling analysis of genetic distance estimates based on the Mahalanobis distance matrix of 12 phenotypic traits measured in 2000.

Table 4. The frequencies of RAPD fragments not present in either *G. max, G. soja*, or semiwild accessions.

	Frequency		
Fragment	Semiwild	*G. max*	*G. soja*
$OPG11_{1900}$†	70% (21)‡	0% (0)	16% (5)
$OPG11_{2500}$	63% (19)	0% (0)	0% (0)
$OPH02_{2100}$	7% (2)	0% (0)	48% (15)
$OPO01_{700}$	0% (0)	32% (10)	0% (0)
$OPO01_{850}$	0% (0)	0% (0)	16% (5)
$OPO05_{2150}$	60% (18)	0% (0)	6% (2)
$OPX05_{450}$	7% (2)	0% (0)	42% (13)

† Primer designation and approximate molecular weight of specific fragment.

‡ Number of accessions.

by Thompson and Nelson (1998) for only *G. max* (30%) but only slightly higher than in the Li and Nelson (2001) data for both of *G. max* and *G. soja* (56%). Fragments $OPO01_{700}$, $OPO01_{850}$, $OPX05_{450}$, $OPH02_{2100}$, $OPO05_{2150}$, $OPG11_{2500}$, and $OPG11_{1900}$ were not found in one or more of the three classes, and a unique fragment was found within each class (Table 4). $OPG11_{2500}$ was only found within the semiwild accessions and occurred in a majority of those accessions. $OPG11_{1900}$ and $OPO05_{2150}$ were present in 60% or more of the semiwild accessions, were totally absent from the *G. max* lines, and existed in low frequencies within *G. soja*. Neither the theory that semiwild-types are evolutionary intermediates between *G. max* and *G. soja* nor that they are products of more recent hybridizations provides a good explanation for unique fragments in this class of accessions, especially not for a fragment that occurred in more than 60% of the semiwild lines. Although the accessions in this study represent only a small portion of the available germplasm, this unique RAPD fragment indicates that the semiwild accessions are in some way genetically distinct from the standard types of the two annual species. $OPH02_{2100}$ and $OPX05_{450}$ occurred in over 40% of the *G. soja* lines, but $OPO01_{850}$ was the only fragment that was found only in *G. soja* lines. The low frequency of $OPO01_{850}$ in these *G. soja* accessions may be one explanation for why it did not occur in either of the other groups that are derived from *G. soja*. $OPO01_{700}$ was found only in the *G. max* lines. Li and Nelson (2001) also reported this as a unique band in *G. max*. It is possible that changes in this region of the genome are partially responsible for the evolution of *G. max*. Extensive research would be required to confirm that $OPG11_{2500}$, $OPO01_{700}$, and $OPO01_{850}$ are unique markers for semiwild, *G. max*, and *G. soja*, respectively, but they do demonstrate the genetic separation of these closely related groups. Removing these taxon-specific fragments from the analysis did not change the cluster groupings. The pattern of divergence among the three classes was primarily attributable to differences in fragment frequencies.

To estimate the number of clusters that should be generated on the basis of the RAPD data, we examined the CCC, PSF, and PST^2 statistics from the output of PROC CLUSTER. All three statistics indicated the presence of three clusters. Multidimensional scaling (MDS) (Fig. 2) and principal component analysis (PCA) separated the accessions into groups that generally corresponded to classifications based on phenotypic data. MDS and PCA put G16 with the *G. max* lines and had G06, G75, G37, and G70 closer to the *G. max* accessions than the other semiwild lines. The two procedures were also consistent in classifying G12, G13, and G63 among the *G. soja* accessions, but in the fifth dimension of MDS, G12 could be separated from *G. soja* accessions. G12 and G13 both possessed $OPH02_{2100}$ and $OPX05_{450}$. These fragments were absent in all *G. max* lines, were in fewer than 10% of the semiwild entries but existed in more than 40% of the *G. soja* accessions. On the basis of phenotypic data, these accessions were set apart from the other semiwild accessions but were not associated with the *G. soja* lines.

The Ward's minimum variance and the VARCLUS methods assigned the 92 accessions into three groups. With both procedures, cluster 1 consists of 22 semiwild accessions, cluster 2 has all 31 *G. max* entries plus five semiwild lines (G06, G16, G37, G70, and G75), and cluster 3 contains all 31 *G. soja* entries and three semiwild lines (G12, G13, and G63). Three (G06, G16, and G75) of the five semiwild lines in cluster 2 were also clustered with the *G. max* group by means of the phenotypic data. G37 and G70 have some *G. max* characteristics with moderately large 100-seed weights (6.5 and 8.4 g), intermediate oil concentration (170 mg g^{-1}), high oleic acid concentrations (200 and 185 mg g^{-1}) and low linolenic acid concentration (100 mg g^{-1}). Although G12, G13, and G63 have severe shattering (5), their other phenotypic characters are not typical of *G. soja* accessions. With the VARCLUS procedure, M92, a *G. max* line, clustered in the *G. soja* group. However, M92 is phenotypically much more like a *G. max* accession with a large 100-seed weight (13 g), high oil concentration (188 mg g^{-1}), high oleic acid concentration (273 mg g^{-1}), low linolenic acid concentration (59 mg g^{-1}), little shattering (2), and intermediate lodging (3). DNA was reextracted from M92 and 20 of the most polymorphic primers were retested. These results confirmed the original data.

The semiwild group has the smallest within-cluster genetic distance (0.107), whereas the *G. soja* group has the largest genetic distance (0.219). These results agree with Maughan et al. (1995) and Li and Nelson (2001) showing the greatest genetic diversity in *G. soja.* The genetic distance between the semiwild cluster and the *G. max* cluster (0.199) was the least distance among the clusters indicating that the semiwild accessions have a closer relationship to *G. max* than to *G. soja.* Broich and Palmer (1980) also showed the semiwild and the *G. max* to be more closely related than either was to *G. soja*. If the semiwild-types are evolutionary intermediates between *G. soja* to *G. max*, theoretically semiwild-types should have a greater genetic variation than *G. max* and less than *G. soja.* If the semiwild accessions are hybridization products presumably only a small proportion of the plants from *G. soja* and *G. max* would have hybridized, which would cause the semiwild-type to have a narrower genetic base than either of the paren-

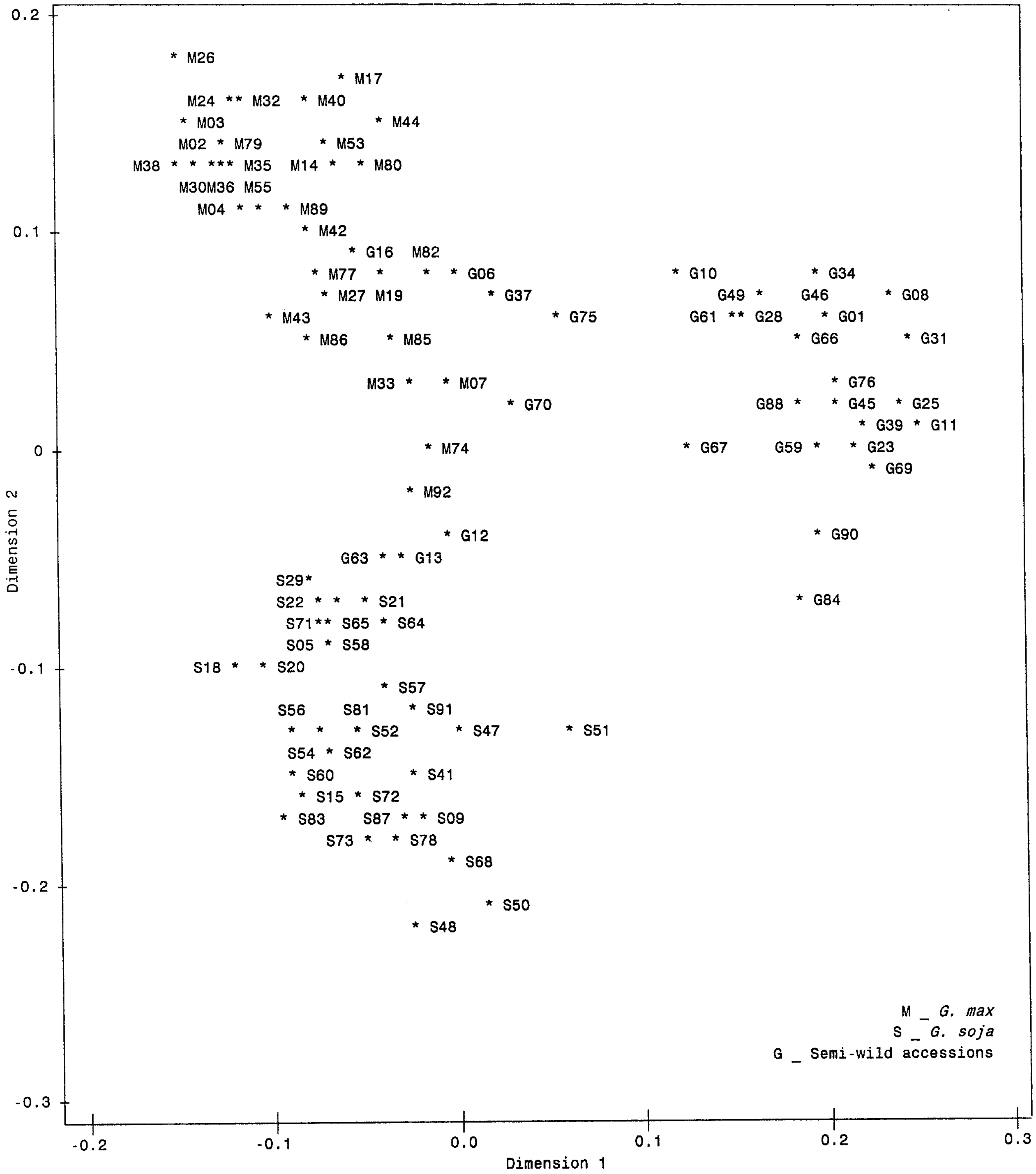

Fig. 2. Two dimensional scatter plot of 92 accessions of *G. max*, *G. soja*, and semiwild accessions obtained from multidimensional scaling analysis of genetic distance estimates based on Jaccard's genetic dissimilarity matrix of 231 RAPD fragments generated by 44 primers.

tal gene pools. If these assumptions are true, the data from this research support the theory that semiwild accessions are hybridization products.

The origin information for many of the *G. soja* lines is more precise than for the other accessions in this study. S50, S60, S72, and S73 were collected from the same field in Shenyang, China (41.48°N) (Bernard et al. 1989). The genetic distances among these four accessions range from 0.096 to 0.186. S22 and S87 were collected in the same pasture near Gongzhuling, Jilin, China (43.32°N) (Bernard et al., 1989). They were phenotypically similar but the RAPD profiles were quite

Table 5. Phenotypic means for *G. max*, *G. soja*, and semiwild accessions inconsistently assigned to taxonomic classes by phenotypic data and DNA markers.

Accession	G12	G13	G63	M92	G06	G16	G75	G37	G70
Classified by phenotypic data	Semiwild	Semiwild	Semiwild	*G. max*	*G. max*	*G. max*	*G. max*	Semiwild	*G. max*
Classified by RAPD data	*G. soja*	*G. soja*	*G. soja*	*G. soja*†	*G. max*	*G. max*	*G. max*	*G. max*	*G. max*
Phenotypic traits									
Lodging (score of 1 to 5)	4	4	4	3	4	4	4	3	4
Shattering (score of 1 to 5)	5	5	5	2	3	5	4	5	3
Stem ratio‡	7.4	8.2	8.4	4.9	8.2	5.6	9.7	7.9	7.4
Seed weight (g 100 seeds^{-1})	2.9	3.1	4.1	13.1	8.9	8.4	8.3	6.5	8.4
Protein (mg g^{-1})	403	457	436	426	404	437	39	387	386
Oil (mg g^{-1})	136	144	157	188	195	177	191	172	173
Oleic acid (mg g^{-1})	192	190	162	273	222	231	192	203	185
Linolenic acid (mg g^{-1})	95	107	111	59	82	89	103	102	97

† M92 was associated with the *G. soja* group only with the VARCLUS analysis.

‡ Ratio of stem diameter for the first internode and the last internode on the main stem.

different. The genetic distance between these two lines was 0.228, which is higher than the within *G. soja* cluster mean genetic distance (0.219). These data demonstrate that *G. soja* lines collected from the same area can be similar but in some cases are genetically quite distinct. It may be necessary to collect multiple samples within *G. soja* populations to sample completely the genetic diversity.

In this study, the three clusters defined by phenotypic data and DNA profiles are highly consistent and strongly correspond to the original *G. soja, G. max*, and semiwild classifications. The current methods for defining *G. max*, *G. soja*, and semiwild are effective, and both phenotypic and DNA marker data can be used to classify *Glycine* accessions. Some accessions (G12, G13, G63, M92, G06, G16, G75, G37, and G70), mostly semiwild, were not consistently classified (Table 5). Four semiwild accessions (G06, G16, G70, and G75) were grouped with *G. max* on the basis of both phenotypic and genotypic data; three semiwild (G12, G13, and G63) and one *G. max* (M92) lines were clustered with *G. soja*; and one semiwild line (G37) was grouped with *G. max* on the basis of genotypic data. Although the three semiwild lines (G12, G13, and G63) share some attributes with *G. soja*, these accessions do not have the plant type that would justify changing their species classification. The five semiwild lines that were grouped with *G. max* are on that ambiguous boundary between semiwild and cultivated and could be included in either group. The results from this study showed that no single trait can be used to distinguish the semiwild soybean. Several characteristics seem to be unique for *G. soja,* and phenotypically seed weight and stem characteristics can be definitive. On the basis of classical taxonomy and cytogenetics, most authors have supported removing *G. gracilis* from the species rank and incorporating it into *G. max* (Hermann, 1962; Wang, 1976; Singh and Hymowitz, 1989).

The results from this study show that semiwild accessions can be distinguished from *G. max* and *G. soja* on the basis of phenotype and DNA markers, but do not necessarily support a separate species designation. These data do help to clarify better the diversity that exists within the annual *Glycine* germplasm and will provide useful information for establishing a core collection of annual *Glycine*.

ACKNOWLEDGMENTS

The authors thank Dr. Richard F. Wilson for providing the chemical composition data used in this research.

REFERENCES

Bernard, R.L., G.A. Juvik, and R.L. Nelson. 1989. USDA Soybean Germplasm Collection Inventory. Vol. 2. International Agricultural Publications. INTSOY Series Number 31. INTSOY, University of Illinois at Urbana-Champaign.

Broich, S.L., and R.G. Palmer. 1980. A cluster analysis of wild and domesticated soybean phenotypes. Euphytica 29:23–32.

Broich, S.L., and R.G. Palmer. 1981. Evolutionary studies of the soybean: The frequency and distribution of alleles among collections of *Glycine max* and *G. soja* of various origin. Euphytica 30: 55–64.

Chang, R.Z., L.J. Qiu, J.Y. Sun, Y. Chen, X.H. Li, and Z.Y. Xu. 1999. Collection and conservation of soybean germplasm in China. p. 172–176. *In* Proc. World Soybean Research Conference VI. Chicago, IL. 4–7 August 1999. National Soybean Research Lab., Urbana, IL.

Dae, H.P., K.M. Shim, Y.S. Lee, W.S. Ahn, J.H. Kang, and N.S. Kim. 1995. Evaluation of genetic diversity among the *Glycine* species using isozymes and RAPD. Korean J. Genet. 17:157–168.

Dong, Y.S., H. Sun, B. Zhuang, L. Zhao, and M. He. 1999. The genetic diversity in annual wild soybean. p. 147–155. *In* Proc. World Soybean Research Conference VI. Chicago, IL. 4–7 August 1999. National Soybean Research Lab., Urbana, IL.

Doyle, J.J. 1988. 5S ribosomal gene variation in soybean (*Glycine*) and its progenitor. Theor. Appl. Genet. 75:621–624.

Doyle, J.J., and R.N. Beachy. 1985. Ribosomal gene variation in soybean (*Glycine*) and its wild relatives. Theor. Appl. Genet. 70: 369–376.

Fei, D.W., and S.Y. Chen. 1996. Reestablish the relationships of species of *Glycine* Genus by RAPD. J. Hered. 23(6):460–468.

Fukuda, Y. 1933. Cytogenetical studies on wild and cultivated Manchurian soybean (*Glycine* L.). Jpn. J. Bot. 6:489–506.

Gizlice, Z., T.E. Carter, Jr., and J.W. Burton. 1996. Genetic diversity patterns in North American public soybean cultivars based on coefficient of parentage. Crop Sci. 36:753–765.

Hermann, F.J. 1962. A revision of genus *Glycine* and its immediate allies. USDA Tech. Bull. 1268. U.S. Gov. Print. Office, Washington, DC.

Hymowitz, T. 1970. On the domestication of the soybean. Econ. Bot. 23:408–421.

Hymowitz, T., and R.J. Singh. 1987. Taxonomy and speciation. p. 23–48. *In* J.R. Wilcox (ed.) Soybeans: Improvement, production, and uses. Agron. Monogr. 16. 2nd ed. ASA, Madison, WI.

Keen, N.L., R.L. Lyne, and T. Hymowitz. 1986. Phytoalexin production as a chemosystematic parameter with *Glycine* ssp. Biochem. Syst. Ecol. 14:481–486.

Kisha, T., C.H. Sneller, and B.W. Diers. 1997. Relationship between genetic distance among parents and genetic variance in populations of soybean. Crop Sci. 37:1317–1325.

Kollipara, K.P., R.J. Singh, and T. Hymowitz. 1997. Phylogenetic and genomic relationships in the genus *Glycine* Willd. Based on sequences from the ITS region of nuclear rDNA. Genome 40:57–68.

Kresovich, S., W.F. Lamboy, R. Li, A.K. Szewc-McFadden, and S.M. Bliek. 1994. Application of molecular methods and statistical analysis for discrimination of accessions and clones of vetiver grass. Crop Sci. 34:805–809.

Li, Z., and R.L. Nelson. 2001. RAPD marker diversity among soybean and wild soybean accessions from four Chinese provinces. Crop Sci. 41:1337–1347.

Maughan, P.J., M.A. Saghai Maroof, and G.R. Buss. 1995. Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. Genome 38:715–723.

Palmer, R.G., K.E. Newhouse, R.A. Graybosch, and X. Delannay. 1987. Chromosome structure of wild soybean (*Glycine soja* Sieb. & Zucc.) accessions from China and the Soviet Union. J. Hered. 78:243–247.

SAS Institute. 1999. SAS/STAT user's guide, Version 8.0, First ed., SAS Inst., Inc., Cary, NC.

Shepard, R.N. 1974. Representation of structure in similarity data: Problems and prospects. Psycometrika 39:373–421.

Shoemaker, R.C., P.M. Hatfield, R.G. Palmer, and A.A. Atherly. 1986. Chloroplast DNA variation and evolution in the genus *Glycine* subgenus *Soja*. J. Hered. 77:26–30.

Singh, R.J., and T. Hymowitz. 1989. The genomic relationships between *G. soja* Sieb. and Zucc. *G. max* (L.) Merr. and '*G. gracilis*' Skvortz. Plant Breed. 103:171–173.

Skvortzow, B.V. 1927. The soybean-wild and cultivated in Eastern Asia. Proc. Manchurian Res. Soc. Pub. Ser. A. Nat. History Sec. 22:1–8.

Smartt, J. 1984. Gene pools in grain legumes. Econ. Bot. 38:24–35.

Thompson, J.A., and R.L. Nelson. 1998. Core set of primers to evaluate genetic diversity in soybean. Crop Sci. 38:1356–1362.

Thompson, J.A., R.L. Nelson, and L.O. Vodkin. 1998. Identification of diverse soybean germplasm using RAPD markers. Crop Sci. 38:1348–1355.

Wang, C.L. 1976. Review on the classification of soybean (in Chinese). Acta Phytotaxon. Sin. 14:22–30.